1996 CENSUS OF CANADA

IAC

Introduction to Census and Automated Coding

Automated Coding

Statistics Canada Statistique Canada

Canada

IAC

Introduction to Census and Automated Coding

Prepared by: Census Operations Division
Social, Institutions and
Labour Statistics Field

Table of Contents

I. Introduction

This manual will introduce you to the 1996 Census of Population and Automated Coding (AC).

II. Overview of the Census

A. General

The Census of Population is Canada's largest and most comprehensive survey. The census collects information on every man, woman and child living in Canada. It can be best described as an official count or a national inventory of all Canadians.

Historically, the first census in Canada, as a measure of social and economic progress, originated under the regime of Jean Talon. In 1666, Jean Talon listed 3,215 persons by age, sex, marital status and occupation.

There were many censuses taken in Canada after 1666, but the first national decennial census was not taken until 1871. Under the *British North America Act, 1867,* the taking of a census every 10 years thereafter became a constitutional requirement of the federal government.

With the opening of the West, new settlers poured onto the Prairies. To chart this lightning population growth, quinquennial censuses (every five years) were taken by the Prairie provinces.

Following the Western example, the federal government decided that a quinquennial census would prove useful as a mid-decade statistical measurement.

With the passage of the new *Statistics Act* in 1971, the Canadian government now takes a census every five years. Canada needs a census every five years to accurately reflect the changes taking place in society. It also allows data users to make comparisons over time. Census questions have been carefully designed and tested to ensure that they can be answered readily by the vast majority of households. The accuracy of census results depends on the good will of Canadians in responding truthfully and accurately to the census questions. Since accuracy depends on having complete information, Parliament provides that the census be mandatory. At the same time, Parliament has enacted confidentiality provisions in the *Statistics Act*, which prohibit the disclosure of any information which could identify a person.

Census-taking, like society, has changed a great deal since Talon's era. To accommodate the increased demand for socio-economic information, self-enumeration, sophisticated processing and coding operations, and computer storage of data have replaced Talon's head count method.

A technique called sampling, whereby only some households are required to answer the long questionnaire (2B), is used in the census because it is cost-effective and reduces the burden on respondents. The long questionnaire is distributed to a sample of one in five households or 20% of the population; the remaining 80% of the population receives the short questionnaire (2A). The short questionnaire provides information necessary for basic decision making. It includes seven questions relating to age, sex, marital status, common-law unions, and first language learned in childhood. The long questionnaire contains the same questions as the short questionnaire, as well as additional questions on topics such as income, education, ethnic origin, labour force activity, mobility, language, and housing for a total of 48 questions. These additional questions provide important information which are used for the benefit of all Canadians.

The results of the census are used in making decisions about our neighbourhoods, communities, provinces and country. Governments, businesses, community groups, healthcare providers, medical researchers, and organizations throughout the country use census data to deal with issues from human resource policies to local education, training, health promotion, and community support programs and services. For example, census data helps to identify future employment needs to aid in the planning of education and training programs. Also, with the help of census data, decision makers can determine the need for roads, schools, day care centres, public transit, and job programs for young people. The census is also needed to support a number of laws and statutes.

The census is the main source of statistics in the country. It is used as a base for calculating ratios, indices, rates, and more. Each person counted results in annual transfer payments from the federal government to the provinces. A province loses hundreds of dollars for each person not counted in the census. In the same way, provinces and territories make grants to local and municipal governments calculated on their total population counts.

The first release of census data is available six months after the close of field work. During that six months, more than 12.4 million questionnaires go through data entry processing, automated coding and compilation before the resulting data can be analysed, printed and made available to the public. Each step must be completed and certified before the data are released. This ensures that census data continue to meet the high standards our users have come to expect. Census data can be obtained from 52 depository libraries across Canada and from bookstores carrying government publications. As well, census information is available from the Advisory Services areas of Statistics Canada regional offices.

B. Census Boundaries

To take a census for a country as large as Canada, certain smaller geographic boundaries must be established in order to facilitate enumeration. The basic boundaries respected are the provinces (PROVs), the federal electoral districts (FEDs) and finally, a smaller unit called the enumeration area (EA). A verification number (VN) is a check digit used to verify the above PROV/FED/EA identification number.

Provincial codes are:

Newfoundland	=	10	Manitoba	=	46
P.E.I.	=	11	Saskatchewan		=47
Nova Scotia	=	12	Alberta	=	48
New Brunswick	=	13	B.C.	=	59
Quebec	=	24	Yukon Territory	=	60
Ontario	=	35	Northwest Territories	=	61

A FED is a geographic area defined by an act of Parliament. The census uses it for field administrative purposes. Every FED elects a member to the House of Commons. FEDs within each province are assigned 3-digit numeric codes ranging from 001 to 099.

An EA is a geographic area for which a census representative is responsible. For purposes of identification, each EA corresponds to a specific numerical code. EAs within FEDs are also assigned 3-digit numeric codes ranging from 001 to 999. Private dwellings (may consist of a family group with or without other non-family persons, of two or more families sharing a dwelling, of a group of unrelated persons or of one person living alone) range from 001 to 799 and collective dwellings (a dwelling of a commercial, institutional or communal nature identified by a sign on the premises or by a census representative speaking with the person in charge or with a resident or a neighbour) range from 901 to 999. Consequently, each geographic area enumerated by one census representative (CR) will be assigned a unique multi-digit identification code. For example, the 9-digit ID number 10-006-036-7 (= PROV/FED/EA/VN) represents:

10	Newfoundland;
006	6th FED identified;
036	36th EA identified;
7	verification number.

There are approximately 52,000 EAs across Canada and each is assigned a unique ID number (PROV/FED/EA/VN).

C. Census Phases

The Census of Population is a very complex undertaking, as it takes about seven to eight years from inception to completion, and involves many groups of employees who are drawn from various divisions throughout Statistics Canada. The initial planning and preparation for a census are set in motion four years in advance of the census year in order to determine what is to be done and to plan the financial and human resources needed for this project.

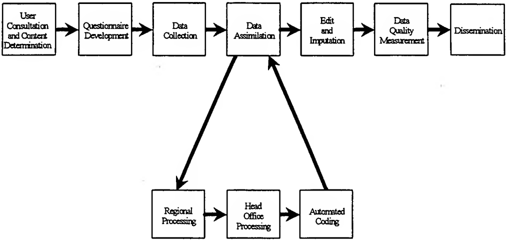
Planning a census begins with a strategic planning exercise, which questions the census content and the process by which the census is taken. Every aspect of the census process is a candidate for change. Changes depend on the evaluation of the previous census, the legislation and policies affecting the census, and the needs and expectations of the various public groups which have an interest in the data produced from the census.

The census process involves a series of automated and manual systems and procedures. They include developing the census questionnaire content and design, delivering the census questionnaires to respondents, processing the data collected, measuring data quality, and disseminating the results.

The 1996 Census of Canada is comprised of seven major phases. These are:

- User Consultation and Content Determination
- Questionnaire Development
- Data Collection
- Data Assimilation
- Edit and Imputation
- Data Quality Measurement
- Dissemination

1996 Census Process Flow Diagram



1. User Consultation and Content Determination

   Statistics Canada asks census users what kind of information they need most. This can include information on the labour market, housing or basic demographic data. Consultation usually begins several years before Census Day.

2. Questionnaire Development

   This portion of the census cycle determines the content and final format of the census questionnaires, the layout, the preparation of artwork, the printing, the packaging and the shipping of the questionnaires to the regional offices.

3. Data Collection

   The field collection process, conducted by census representatives (CRs) in 52,000 enumeration areas (EAs), consists of the drop-off and mail-back of approximately 12,250,000 questionnaires (9,800,000 2A questionnaires, 2,100,000 2B questionnaires, 1,750 2C questionnaires, 133,000 2D questionnaires and 200,000 Form 3 questionnaires) to and from Canadian households. These questionnaires are then edited to ensure they have been properly completed by all Canadians across the country. This phase employs over 45,000 people.

4. Data Assimilation (DA)

   The assimilation process turns the questionnaire responses into machine readable information. During the processing phase of the census cycle, the questionnaire responses go through the following three production operations.

   - Regional Processing (RP)

     RP is defined as the manual preparation and key entry of data from all census questionnaires into a machine-readable format. The manual preparation consists of four operations which are: receipt and registration, coding of economic variables (Industry and Occupation), batching and labelling and finally, shipping. The means of capture is through a key entry system at Revenue Canada (RC). All RP activities are conducted by RC in seven regional centres (St. John's, Jonquière, Shawinigan, Ottawa, Sudbury, Winnipeg, and Surrey). RP activities employ approximately 1,900 people: 650 people for the pre-capture activities and 1,250 for data capture activities.

   - Head Office Processing (HOP)

     HOP receives, controls and stores questionnaires, visitation records (VRs) and data cartridges. HOP is a combination of manual and automated processing designed to carry out structural edits. The processing of questionnaires from Canadians outside Canada (2Cs), merchant, coast guard, and naval vessels is also carried out. HOP sends the census data to Automated Coding (AC) and Edit and Imputation (E&I) and prepares a data file of final population counts for dissemination.

- Automated Coding (AC)

  AC utilizes batch and interactive processing where written responses key entered from the census questionnaires are matched against reference files containing a series of words or phrases and the corresponding numeric codes.

5. Edit and Imputation (E&I)

   In this phase, a system creates a usable database from the information provided by Data Assimilation. E&I checks for blank, invalid and illogical situations and assigns respondent data for subsequent retrieval and publication.

6. Data Quality Measurement

   The objective of this process is to develop and conduct studies on the quality of the census data so users have measures (from quality control operations) or indicators as to the reliability of census results. Studies carried out include Reverse Record Check (RRC), Vacancy Check (VC) and the Coverage Research Study.

7. Dissemination

   This final phase of the census consists of the retrieval and tabulation of census data for publication and custom tabulations. Information is made available in a number of formats, including CD-ROM and diskettes. Dissemination is responsible for developing products and services in response to the needs of the data user community.

III. Automated Coding (AC)

A. Process

The automated coding process includes a combination of automated and manual steps, beginning with the receipt of the written responses from Head Office Processing. Records containing alphanumeric responses, as well as other specific person or household data which may be needed to assist in the numeric coding of those responses, are provided.

Of the 48 questions on the 2B questionnaire, 15 questions (variables) will have responses which will be sent to Automated Coding. Of these 15 variables, three major groups are formed. They are as follows:

- Relationship to Person 1;
- Sociocultural variables:
  - Mother Tongue;
  - Home Language;
  - Non-official Language;
  - Place of Birth;
  - Indian Band/First nation;
  - Ethnic Origin;
  - Place of Residence 1 Year and 5 Years Ago Inside Canada;
  - Place of Residence 1 Year and 5 Years Ago Outside Canada;
  - Major Field of Study;
  - Citizenship;
  - Population Group;
- Place of Work.

The reason for the formation of these three major groups are the systems developed to code these variables in Automated Coding.

A mainframe-based system was developed to code the sociocultural variables. This system was initially developed to code sociocultural variables during the 1991 automated coding process and was refined for use during the 1996 automated coding process.

Two separate PC-based systems (FOXPRO) were developed to code both the Relationship to Person 1 and Place of Work variables. Both these systems have been developed especially for the 1996 automated coding process.

The primary objective within all three coding systems is to match written responses to predetermined reference files which in turn assigns a numeric code to the written response.

For example: If a respondent writes "Algeria" as his Place of Birth, the system will search through the reference file associated with the variable Place of Birth and will look for the word "Algeria". When the system locates the word "Algeria", it will assign it a corresponding numeric code, for example, 647.

Written Response		Reference File	
Algeria	no	Afghanistan	701
	no	Albania	566
	yes	Algeria	647
		American Samoa	801

If the system cannot locate a word within the reference file (for example, if the word is not spelt correctly by the respondent or if the word is abbreviated), then the system will send this response to the next step in the automated coding process which is called interactive coding.

During interactive coding, coders will review written responses which could not be coded by the system and attempt to assign them a numeric code using predetermined procedures.

Once all responses for a variable have been coded, either by the system or manually, a sample is passed through quality control, and a quality control evaluation is completed. The resulting codes are then stored for later transfer to Edit and Imputation in order to continue with census processing.

The three systems developed also have their own distinct characteristics. The following paragraphs outline some of these differences between each of the three coding systems.

- The Relationship to Person 1 coding system

  Responses are coded for two types of households: private and collective. A private dwelling may consist of a family group with or without other non-family persons, of two or more families sharing a dwelling, of a group of unrelated persons or of one person living alone. A collective dwelling refers to a dwelling of a commercial, institutional or communal nature identified by a sign on the premises or by a census representative speaking with the person in charge or with a resident or a neighbour and who do not have a usual place of residence elsewhere in Canada.

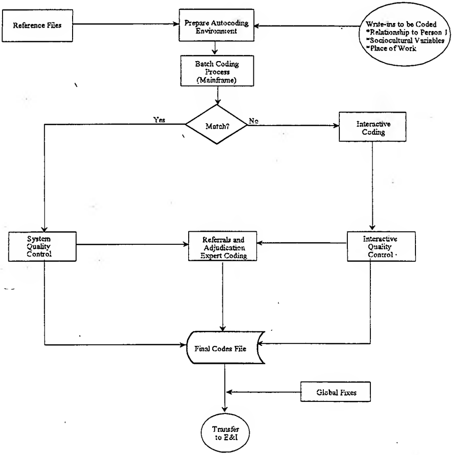
- The sociocultural variables coding system

  The sociocultural variables coding system is based on a mainframe computer software. Both processes, i.e. matching and interactive coding are done on mainframe terminals.

- The Place of Work coding system

  The major difference between the Place of Work coding system and the two other systems developed for the AC process is that this one uses five reference files to match written responses instead of one. The system will look through five reference files to locate a "Place of Work" before it decides whether it can match a response or send it to interactive coding. Interactive coders also have a choice of five reference files when coding Place of Work (POW) response. The code assigned during POW coding is a geographic location.

1996 AUTOMATED CODING
PROCESS FLOW



B. Coding Activities

All three interactive coding systems use different levels of coders also referred to as tiers of coders. The Relationship to Person 1 and sociocultural variables are coded by two levels of coders, Tier 1 and Tier 2 coders also referred to as general coders and expert coders. The Place of Work variable is coded by three levels of coders, Tier 1, Tier 2 and Tier 3 also referred to as general, referral, and expert coders.

General coders (Relationship to Person 1, sociocultural variables and Place of Work) manually code misspelled responses, abbreviated responses, specific cases such as "Indian, Native Canadian, Same", multiple responses (a multiple response occurs when a respondent has entered two or more responses for a single question on the questionnaire) and other cases. They also perform quality control coding of codes assigned by the system or by coders. Responses which general coders are unable to code will be referred to the next level of coding.

Referral coders (Place of Work) code responses referred by general coders. They have access to additional reference files available only to referral and expert coders to help them with coding. Referral coders also adjudicate (recode) Place of Work code/response combinations on which general coders and first quality control disagree. Referral coders recode entire "lots" of general coded responses which failed first quality control. Lastly, referral coders assign a Place of Work code to those responses which were coded by other referral coders and sampled for second quality control.

Expert coders (Relationship to Person 1, sociocultural variables and Place of Work) perform the same coding options as general and referral coders. They also perform the following activities:

- they code responses referred to a more expert level of coding;
- they code responses deferred to a later time. They resolve responses deferred to a later time during a previous coding session. These responses have been marked as needing more time for research;
- they recode rejects. They review and recode responses from rejected work units coded by coders of the previous level. Responses from a rejected work unit to be recoded are those which were not part of the quality control sample;

- they perform quality control. They review sampled responses, which as a result of the second quality control coding are majority referrals or three-way discrepancies. The code assigned by the expert coder is compared to the production code, the QC1 code and the QC2 code. Each code which does not agree with the final code assigned by the expert is attributed an error;
- they are authorized to assign a final code of "Unable to code". The code assigned by an expert coder is final and is not subject to quality control;
- they maintain and update the reference files.

IV. System and Interactive Quality Control Coding

Quality control procedures are performed on both the system matched and the interactively coded responses to ensure that an acceptable level of quality is maintained. Quality Control assists in detecting operational problems in codes assigned by the system and interactively.

First quality control coding consists, for a coder, in assigning a code to a sampled response previously coded by either the system or by another coder. The two codes are compared by the system. If they are the same, the code is accepted and considered final. If there is a discrepancy between the two codes (two-way discrepancy), the response is sent to second quality control coder where he/she assigns a code to the sampled response. These three codes are then compared by the system. If two of the codes are the same, this code is considered correct and final and an error is attributed to the discrepant code (majority rule). If none agree (three-way discrepancy), the response is sent to expert quality control coding for review by an expert coder (majority referral).

The general or referral coder will not know whether the response on the screen has been previously coded or not. All coding errors made, either as an original general coder or as a general quality control coder, are recorded and monitored.

The subject-matter specialist responsible for each variable reviews the results and monitors levels of error.

V. Conclusion

Now that you have read the Introduction to Census and Automated Coding Manual, you are ready to proceed to the next section of this training session, in which you will learn how to code responses for specific variables. If you have any remaining questions, be sure to ask your supervisor for clarification.